Chapter 134
Sentiment Analysis of Text Using SVM

Yong Yang, Chun Xu and Ge Ren

Abstract In recent years, development with the internet, information processing turns more and more important for us to get useful information. Text Categorization, the automated assigning of natural language texts to predefined categories based on their contents, is a task of increasing importance. Therefore, an approach of text emotional classification based on support vector machine was proposed. The system includes four parts: word segmentation, establishment emotional word database, training the model of classification, and testing. The system is actually used for classification of the reader's emotions that classify the text to achieve the purpose of forecast readers' emotion.

Keywords Text · Sentiment analysis · Classification · Support vector machine · Word segmentation

Y. Yang (✉)
Xinjiang Technical Institute of Physics and Chemistry, Chinese
Academy of Science, Urumqi, China
e-mail: Yangyong1900@163.com

Y. Yang · G. Ren
College of Computer Science, XinJiang Normal University, Urumqi, China
e-mail: RenGe@163.com

C. Xu
College of Computer Science and Technology, Xinjiang University of
Finance and Economics, Urumqi, China
e-mail: XuChun@163.com

X. Wang et al. (eds.), *Electrical, Information Engineering and Mechatronics 2011*,
Lecture Notes in Electrical Engineering 138, DOI: 10.1007/978-1-4471-2467-2_134,

134.1 Introduction

The booming Internet has spread its influence on people throughout all aspects of social life. In recent years, online communities, blogs, and multiple kinds of forums has offered people a broader platform for information exchange with the conversion of most Internet users from passively receiving information to initiatively creating information. The large number of Internet information has expressed people's emotional coloring as well as their tendencies, such as "Happy, Angry, Sorrow, and Joy". In-depth popular opinions on certain incident or product can be obtained through emotional analysis of text, whereas the rapid expansion of information amount has made it impossible to accomplish text emotional analysis merely by manual methods, for which sake text emotional analysis in terms of computer technology has been introduced [1].

The second part of this article elaborates on the methods adopted in this article during emotional information extraction phase, which include measurement algorithm (TF-IDF) and classification algorithm Support Vector Machine (SVM); the third part details the methods specific to the accomplishment of text emotion analysis in this article; the fourth part lists the specific experiment as well as its results; the fifth part comes the conclusion of this article.

134.2 Algorithm Description

134.2.1 SVM Classification Algorithm

Training methods and classification algorithms are core section in classification system. Currently there are a variety of training algorithms and classification algorithms based on vector space model, such as SVM, neural networks, the average maximum entropy method, close to recent methods, and Bayesian K, etc. [2]. This article takes SVM algorithm during the specific system implementation.

Establish the sample set $(x_i, y_i),\ i = 1, \ldots, n\, y \in \{-1, +1\}$ is the category label. The general form of linear discriminant function as well as separating hyperplane equation in d-dimensional space is expressed as $g(x) = w \cdot x + b$ and $w \cdot x + b = 0$, normalize the discriminant function to ensure all samples meet $|g(x)| \geq 1$ and the nearest sample to separate hyperplane meeting $|g(x)| = 1$ with the classification interval equal to $2/||w||$ and make sure that the separating intervals are maximum equivalent to minimizing $\|w\| \left(\text{or } \|w\|^2\right)$ Requiring the classification hyper-plane to correctly classify all samples means that the classification hyper-plane should meet:

$$y_i[(w * x_i) + b] - 1 \geq 0, i = 1, \ldots n \tag{134.1}$$

Therefore, the hyperplane which meets formula (134.1) while minimizing $\|w\|^2$ value is called optimal separating hyperplane, and the training samples on H1, H2 corresponding to the optimal separating hyperplane is defined as support vectors. *w**, the solution to *w*, can be gained from optimization theory, likewise, *b**, the solution to will be obtained by using support vector machines, after which optimal classification function is gained

$$f(x) = \operatorname{sgn}\{(w^* \cdot x) + b^*\} \tag{134.2}$$

Substitute the optimal classification function to the input sample type.

134.2.2 Text Description

As for text expression, this article uses vector space model (VSM), the basic idea which is :Each text has some individual properties which are represented by a number of concept words for content expression, whereas each property can be seen as a dimension in concept space, and these individual properties are regarded as text feature item, therefore the text is seen as a collection of such items, which releases the necessity to consider the complex relationships among paragraphs, sentences, and words in text structure Hereby the text can be represented as a form like d = (t1, *W*1;t2, *W*2;……;tn, *W*n), among which ti and *W*i refer to the feature item and correspondent weight [3–5]. The feature item weight *W*i is commonly measured by TF-IDF formula which is defined as follows:

$$W_i = \frac{tf_{ij} \times \log(N/n_i + 0.01)}{\sqrt{\sum_{t \in \check{d}} \left[tf_{ij} \times \log(N/n_i + 0.01)\right]^2}} \tag{134.3}$$

Here tf_{ij} refers to the frequency of feature item *ti* in text *d* with *N* being the total text number, n_i being the number of text in which *ti* has appeared in text set, and the denominator refers to normalizing factor [6].

Obviously, feature items with large weight refer to those words found sufficient frequent in one text while rarely seen in other texts in the whole text set, which will play an important role in text classification.

134.2.3 Feature Extraction

A large number of words are necessary for one text, in response; the dimension of vector space used for text expression is also quite large (may be up to 10,000 above). Therefore, it is mandatory to implement dimension reduction for the following two reasons [7]: 1. Enhancement in program efficiency and operation speed. 2. Each of all the words (may be up to ten thousands above) contributes differently to text classification in that those common words which are available in each category contribute less whereas those found frequent in certain specific category while rarely seen in other categories are able to make substantial contribution, for which sake in each category it is suggested that those less expressive words be removed so as to screen out the feature item set specific to that category for classification accuracy, and feature scoring function constructed based on mutual information, information gain, expected cross entropy, and text evidence is regarded as the common extraction method.

The criterion in which mutual information of words and categories for feature item extraction is considered as one effective method, the basic procedures of which are described as below:

Step 1. The collection of feature items shall contain all words appeared in this category during initial case

Step 2. Calculate the mutual information between words and categories for each word.

$$\log\left(\frac{P\left(W|C_j\right)}{P(W)}\right) \times \exp\left(\frac{N_j}{N}\right) \tag{134.4}$$

In (134.4), $P\left(W|C_j\right)$ refers to the frequency of the feature item W from training corpus in category C_j, $P(W)$ represents the frequency of the feature item in training corpus. N_j is the text number of the feature item W appeared in category C_j, N refers to the text amount for the feature item W appeared in training corpus.

Step 3. Sort all words in this category as per the mutual information amount calculated above.

Step 4. Extract a certain number of words as the feature item.

Finally, 1,000 terms are selected as the feature item in each category.

134.3 Functional Accomplishment

This article has initially accomplished the structure diagram specific to system as per the above mentioned solution to text classification system, which is listed as below Fig. 134.1:

In the pre-processing procedure for training text, text data will undergo words separation, removal of stop words and parameter statistics (including words frequency, relevant text number and weight) in sequence so as to get data represented by text vector model whereas feature item extraction will be realized by using above mentioned mutual information method for words. The training data generation means processing each text into text vector data expressed by feature item and making correspondent adjustment as per the SVM interface specification selected. Pre-processing and data generation of testing text refers to the word separation and expression of feature item vector model process for one unknown text, classification number of the unknown text which is in test will be generated by classification and output with associated statistical accuracy available.

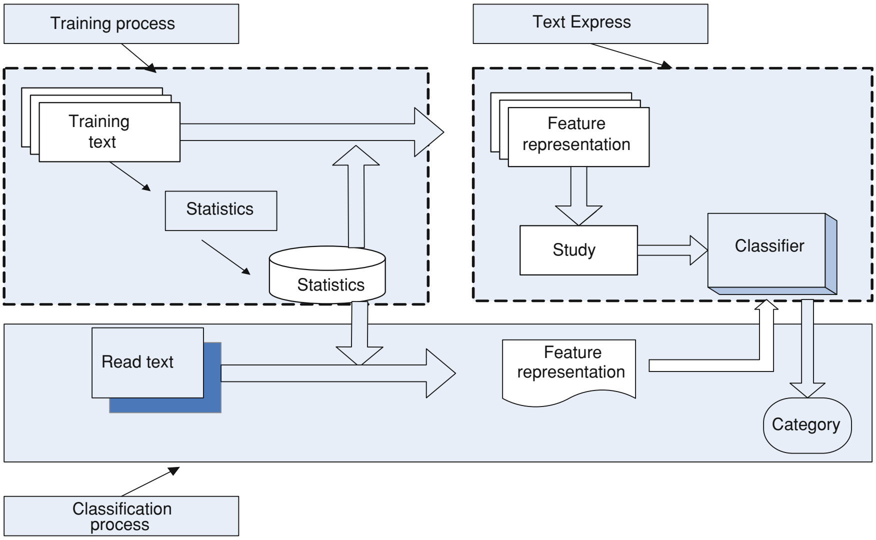

Fig. 134.1 Structure of text classification system

134.4 Analysis and Evaluation on Results

Evaluation index for text theme classification including recall ratio and precision ratio and F value are used in this experiment for emotional classification of text in this article. Precision ratio refers to the percentage of the text that matches the result of manual classification in all tested text; recall ratio is the percentage of the text that matches classification system in all text as per manual classification result. Both precision ratio and recall ration shall be taken into consideration comprehensively since they represent two different aspects in classification quality, therefore one kind of new evaluation index is introduced, which is named as testing value F. Set a1 and a2 as the correct text number in all positive text and negative text determined by classifier respectively; b1 and b2 as the positive text and negative text number determined by classifier respectively; c1 and c2 as positive text and negative text number respectively. It can be obviously seen that $c_1 + c_2 = b_1 + b_2$. The calculation formula for evaluation index is shown in Table 134.1:

Obviously, $F1 = F2 = F$ since $c_1 + c_2 = b_1 + b_2$ All text in corpus are composed of 800 network article and blog posts generally taken from multiple portals and BBS, which are classified into five categories: Funny, Angry, Sorrowful, Novel, and Boring. The method for training set and testing set selection is as below: Divide these classified corpus into five equal sets with one set selected as an open test set and others as training set and closed training set, likewise, each set will be selected as the open training set in turns, run classification algorithm

Table 134.1 Calculation formula for evaluation index

Category measurement	Positive	Negative	Total
Precision ratio	$pp = \frac{a_1}{b_1} \times 100\%$	$pn = \frac{a_2}{b_2} \times 100\%$	$F2 = \frac{a_1 + a_2}{b_1 + b_2} \times 100\%$
Recall ratio	$RP = \frac{a_1}{c_1} \times 100\%$	$RN = \frac{a_2}{c_2} \times 100\%$	$F2 = \frac{a_1 + a_2}{c_1 + c_2} \times 100\%$
F value	$FP = \frac{2 \times RP \times PP}{RP + PP}$	$FN = \frac{2 \times RP \times PP}{RP + PP}$	$F = \frac{2 \times F1 \times F2}{F1 + F2}$

Table 134.2 Experiment results

Category	Recall ratio in closed test (%)	Precision ratio in closed test (%)	F value in closed test (%)	Recall ratio in open test (%)	Precision ratio in open test (%)	F value in open test (%)
Positive	81.08	81.10	81.18	80.33	80.40	80.56
Negative	82.39	81.78	81.08	81.17	81.26	82.11
Total	81.11	81.42	81.25	81.09	81.12	81.20

(total five times classification operations) to get the average value. The experimental results are shown in Table 134.2.

134.5 Conclusion

This article has come up with one solution plan for text emotional system as per SVM classification method, which is based on classification algorithm specific to supervised study model which is different from that to unsupervised study model. This article has introduced the algorithm model for systematic solution and come up with the associated structural plan as well as the implementation method. Meanwhile this article has also described the experiment procedures and the results while conducting further analysis of problems encountered in experiment and proposed recommendations for future work.

Acknowledgment Sponsor acknowledgment: Natural Science Fund of Xinjiang Uygur Autonomous Region. No:2010211A22 Natural Science Fund of China. No: 31040050.

References

1. Yu H, Hatzivassilogou V (2003) Towards answering opinion questions: separating facts from op inions and identifying the polarity of opinion sentences. In: Proceedings of the 2003 conference on empirical methods in natural language processing, Association for Computational Linguistics, Morristown, NJ, pp 129–136
2. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the conference on empirical methods in natural language processing, pp 79–86

3. Dave K, Lawrence S (2003) Mining the peanut gallery: opinion extracts on and semantic classification of product reviews. In: Proceedings of the 12th international conference on knowledge capture. ACM Press, New York, pp 70–77
4. Zhang W, Yoshidab T, Tang X (2010) A comparative study of TF*IDF, LSI and multi-words for text classification [J]. Expert Systems with Applications 38(3):2758–2765
5. Kamps J (2002) Visualizing word net structure. In: Proceeding of the first global word net conference, India, pp 182–186
6. Xu Y, Wang B, Li J (2008) An Extended Document Frequency Metric for Feature Selection in Text Categorization [J]. Lect Notes Compu Sci 4993(20):71–82
7. Pei-yu L, Yu-zhen Y, Jing Z (2011) Text representation combining syntax in vector space model. Adv Inf Sci Serv Sci 3(7):251–259